



Fast solver for band Toeplitz-block-band Toeplitz system

Houssam Khalil

► To cite this version:

| Houssam Khalil. Fast solver for band Toeplitz-block-band Toeplitz system. 2008. hal-00366297v2

HAL Id: hal-00366297

<https://hal.science/hal-00366297v2>

Preprint submitted on 12 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Résolution rapide d'un système de Toeplitz bande par blocs Toeplitz bandes

Houssam Khalil*

Resumé. Soit T une matrice de Toeplitz par blocs de Toeplitz à coefficients dans un corps \mathbb{K} et de taille N . On suppose qu'elle est formée de m blocs de taille $n \times n$; de plus elle est bande par blocs, c'est à dire qu'en dehors des $2k_1 + 1$ diagonales par blocs centrales les blocs sont nuls; les blocs eux-même sont bande : en dehors des $2k_2 + 1$ diagonales centrales, les éléments des blocs sont nuls. On donne dans ce papier trois méthodes de résolution rapide, en $\mathcal{O}(N^{3/2})$, pour résoudre le système linéaire. Ces méthodes sont plus rapides que les méthodes associées aux matrices creuses. On donne aussi une statistique qui montre que la matrice T devient de plus en plus mal conditionnée quand les largeurs des bandes décroissent. Cette remarque n'est pas vrai pour le cas d'une matrice de Toeplitz claire.

1 Introduction

Soit T une matrice de Toeplitz par blocs de Toeplitz à coefficients dans un corps \mathbb{K} . On suppose qu'elle est formée de m blocs de taille $n \times n$; de plus elle est bande par blocs, c'est à dire qu'en dehors des $2k_1 + 1$ diagonales par blocs centrales les blocs sont nuls; les blocs eux-mêmes sont bande : en dehors des $2k_2 + 1$ diagonales centrales, les éléments des blocs sont nuls. T est donc de la forme suivante :

$$T = \begin{pmatrix} T_0 & T_{-1} & \dots & T_{-k_1} & 0 & \dots & 0 \\ T_1 & T_0 & \dots & T_{-k_1+1} & T_{-k_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ T_{k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-k_1} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-1} \\ 0 & \dots & 0 & T_{k_1} & \dots & T_1 & T_0 \end{pmatrix}, \quad (1)$$

*Institut Camille Jordan, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex France (khalil@math.univ-lyon1.fr).

et chaque T_j est de la forme :

$$T_j = \begin{pmatrix} T_{0,j} & T_{-1,j} & \dots & T_{-k_2,j} & 0 & \dots & 0 \\ T_{1,j} & T_{0,j} & \dots & T_{-k_2+1,j} & T_{-k_2,j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ T_{k_1,j} & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-k_2,j} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-1,j} \\ 0 & \dots & 0 & T_{k_2,j} & \dots & T_{1,j} & T_{0,j} \end{pmatrix}. \quad (2)$$

Soit $N = nm$, et on va supposer que m et n sont de même ordre, c'est à dire $C^{-1}m \leq n \leq Cm$ pour une certaine constante $C > 0$. Par conséquent $m = \mathcal{O}(\sqrt{N})$, $n = \mathcal{O}(\sqrt{N})$. Il est classique que la résolution d'un système $n \times n$ de structure bande de largeur k coûte $\mathcal{O}(k^2 n)$ opérations par la méthode de Gauss. Dans notre cas, le coût serait donc en $\mathcal{O}(N^2)$ opérations avec des méthodes directes pour matrices creuses.

Dans ce papier, on va essayer d'exploiter la structure *Toeplitz bande par blocs Toeplitz bande* pour donner une estimation en $\mathcal{O}(N^{3/2})$. On donnera trois algorithmes de résolution rapide en $\mathcal{O}(N^{3/2})$.

Remarquons tout d'abord que des statistiques expérimentales sur le nombre de conditionnement des matrices aléatoires de Toeplitz et des matrices aléatoires de Toeplitz par blocs de Toeplitz montrent des comportements dont il importe de tenir compte dans l'analyse des résultats.

La section (2) décrit ces statistiques ainsi que les moyens utilisés pour les obtenir. Il faut mentionner ici que les comportements dans le cas Toeplitz et dans le cas Toeplitz par blocs de Toeplitz sont qualitativement différents.

2 Statistiques pour des matrices bandes Toeplitz et pour des matrices bandes Toeplitz par blocs bande Toeplitz

2.1 État des connaissances

Un certain nombre de résultats théoriques sont connus pour des matrices aléatoires à structure bande, ou des matrices aléatoires à structure de Toeplitz. Notons en particulier un théorème de limite centrale pour un modèle de matrice bande symétrique par Anderson et Zeitouni [1], un résultat sur la distribution limite des valeurs propres pour les grandes matrices bandes symétriques ou hermitiennes par Molchanov et al. [5], un résultat de grandes déviations pour un modèle de matrices hermitiennes par Guionnet [3], des résultats sur les développements asymptotiques et les échelles d'universalité spectrale par Khorunzhy et Kirsch [4]. D'autre part il y a des résultats pour les matrices de Toeplitz aléatoires, voir [?] et [?].

aucun de ces résultats ne porte sur le nombre de conditionnement de l’une ou l’autre famille de modèles de matrice, alors que c’est l’objet essentiel pour l’analyse numérique des méthodes de résolution de systèmes linéaires.

On a donc décidé d’obtenir des informations expérimentales pour des matrices aléatoires bande Toeplitz et bande Toeplitz par blocs bande Toeplitz et on a procédé comme suit.

2.2 Algorithmes

Notons $2k + 1$ la largeur impaire de bande dans le cas de Toeplitz scalaire et $2k_1 + 1$, $2k_2 + 1$ les deux largeurs impaires de bande dans le cas par blocs. Remarquons que les diagonales non nulles occupent une zone symétrique. Plus précisément, dans le cas scalaire, les diagonales sont nulles en dehors de $\{-k, \dots, k\}$, et dans le cas par blocs, les diagonales de blocs sont nulles en dehors de $\{-k_1, \dots, k_1\}$ et les diagonales dans les blocs sont nulles en dehors de $\{-k_2, \dots, k_2\}$.

On a généré des coefficients pseudo-aléatoires gaussiens en nombre approprié, c’est à dire k dans le cas scalaire et $k_1 k_2$ dans le cas par blocs, ce qui permet de décrire une matrice T en structure creuse.

On a effectué une décomposition LU creuse, par l’algorithme “superLU”, de T et sa transposée, ce qui a permis de calculer la plus petite valeur singulière par la méthode de la puissance inverse. D’autre part on a obtenu la plus grande valeur singulière par la méthode de la puissance.

On a constaté la convergence rapide de la méthode de la puissance inverse, et la convergence très lente de la méthode de la puissance. Ce qui n’est pas entièrement surprenant.

Le rapport entre la plus grande et la plus petite des valeurs singulières fournit le nombre de conditionnement $\kappa(T) = \|T\|_2 \|T^{-1}\|_2$, la norme $\|\cdot\|_2$ étant la norme spectrale, c’est à dire la racine carrée de la plus grande valeur propre de T^*T .

2.3 Les résultats dans le cas scalaire

On se reportera au figure (1) pour voir des histogrammes du logarithme en base 10 du nombre de conditionnement des matrices de Toeplitz de même taille et des largeurs de bande différentes.

On remarque que ces histogrammes dépendent peu de la largeur de bande.

2.4 Les résultats dans le cas par blocs

Dans ce cas on remarque que la statistique sur les nombres de conditionnements dépend des largeurs de bande. On constate qu’avec l’accroissement d’au moins une des largeurs de bande, l’histogramme devient de plus en plus étroit, et la moyenne du logarithme du nombre de conditionnement décroît. Ceci montre que les matrices TBT bande par blocs bande de petites largeurs de bande sont plus mal conditionnées que celles de largeurs de bande plus grandes.

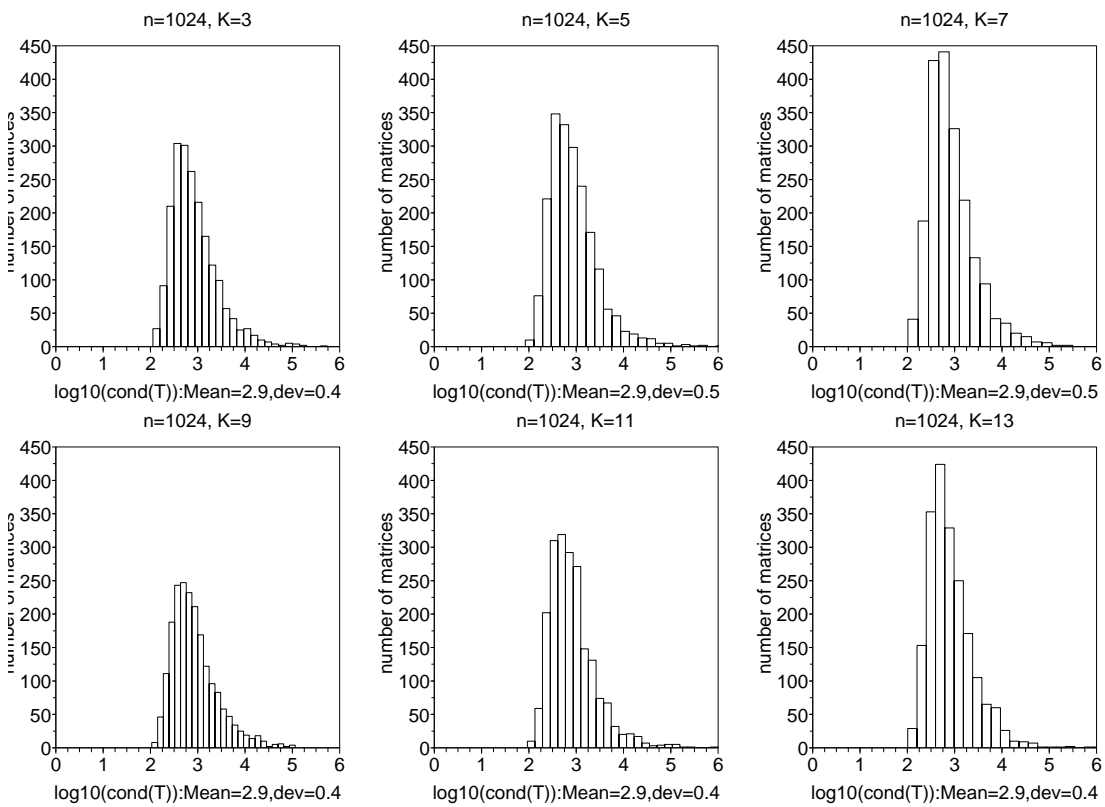


FIG. 1 – Les matrices ici sont de Toeplitz bande de taille 1024×1024 , la largeur de bande, $2k+1$, varie entre 3 et 13. On remarque que le nombre de conditionnement ne varie pas avec la largeur de la bande.

On se reportera aux figures (2), (3), (4) et (5).

Dans la suite, on décrira trois algorithmes rapides de résolution d'une matrice bande en commençant tout d'abord par le cas scalaire, puis en les généralisant au cas par blocs.

3 Cas scalaire

Les idées de deux premiers algorithmes proviennent de Bini et Pan [2], et les auteurs mentionnent un problème d'instabilité pour le premier algorithme.

3.1 Transformation en matrice circulante plus matrice de petit rang

Rappelons, tout d'abord, la Formule de Sherman-Morrison-Woodbury :

Théoreme 1. Soient $A \in \mathbb{K}^{n \times n}$, $G, H \in \mathbb{K}^{n \times k}$. Si $I_k + H^T A^{-1} G$ n'est pas singulière, alors

$$(A + GH^T)^{-1} = A^{-1} - A^{-1}G(I_k + H^T A^{-1}G)^{-1}H^T A^{-1}. \quad (3)$$

Démonstration. Une matrice par blocs peut se factoriser comme suit :

$$L = \begin{pmatrix} M & N \\ P & Q \end{pmatrix} = \begin{pmatrix} I & 0 \\ PM^{-1} & I \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & M^{-1}N \\ 0 & I \end{pmatrix},$$

avec S le complément de Schur donné par $S = Q - PM^{-1}N$. Bien entendu cette factorisation n'a de sens que si M est inversible. Par symétrie, on a également

$$L = \begin{pmatrix} M & N \\ P & Q \end{pmatrix} = \begin{pmatrix} I & NQ^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{S} & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} I & 0 \\ Q^{-1}P & I \end{pmatrix},$$

avec $\tilde{S} = M - NQ^{-1}P$. En inversant les deux termes de l'égalité et en identifiant les deux expressions du bloc de la première ligne et de la première colonne on tire l'identité :

$$M^{-1} + M^{-1}NS^{-1}PM^{-1} = \tilde{S}^{-1}$$

qu'on applique à

$$L = \begin{pmatrix} A & G \\ H^T & -I_k \end{pmatrix}.$$

□

Proposition 1. Soit $A \in \mathbb{K}^{n \times n}$ une matrice de Toeplitz bande, de largeur de bande $2k + 1$. Alors on peut décomposer A sous la forme $C + R$ avec C une matrice circulante et R une matrice de rang au plus $2k$.

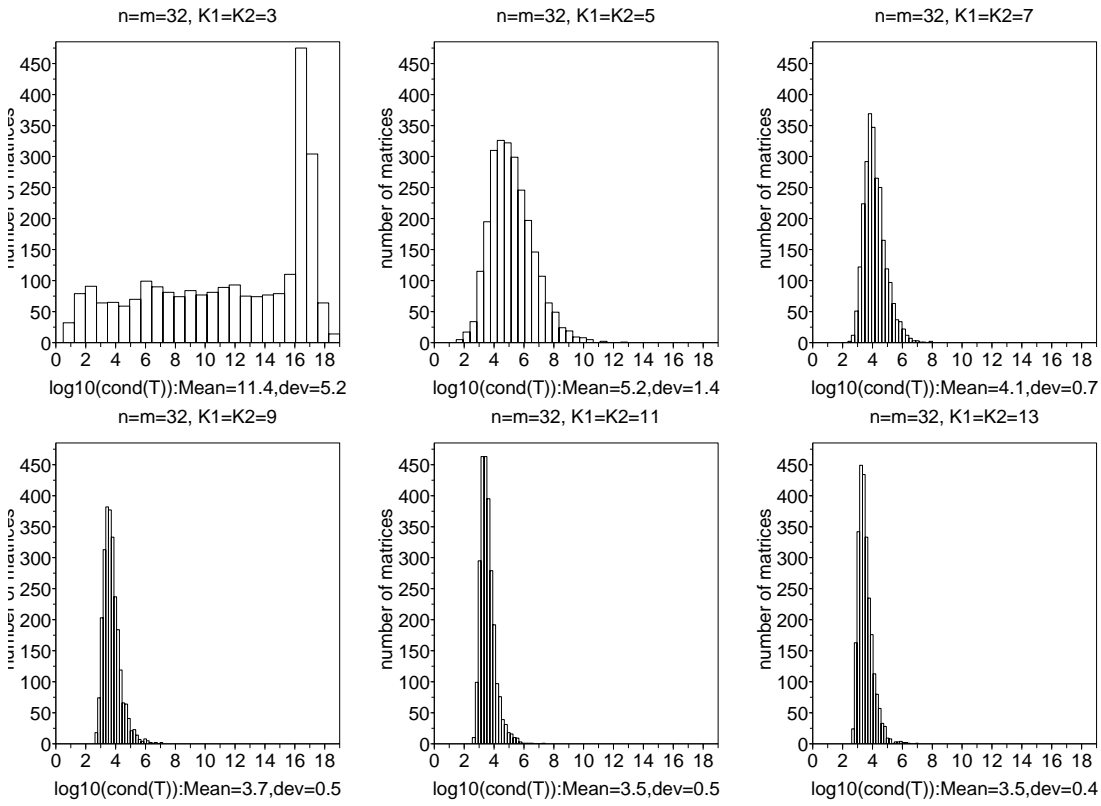


FIG. 2 – On considère des matrices TBT de taille $32^2 \times 32^2$ et de largeur de bande $K_1 = 2k_1 + 1 = 2k_2 + 1$, allant de 3 à 13. On a fait 2500 essais pour chaque cas.

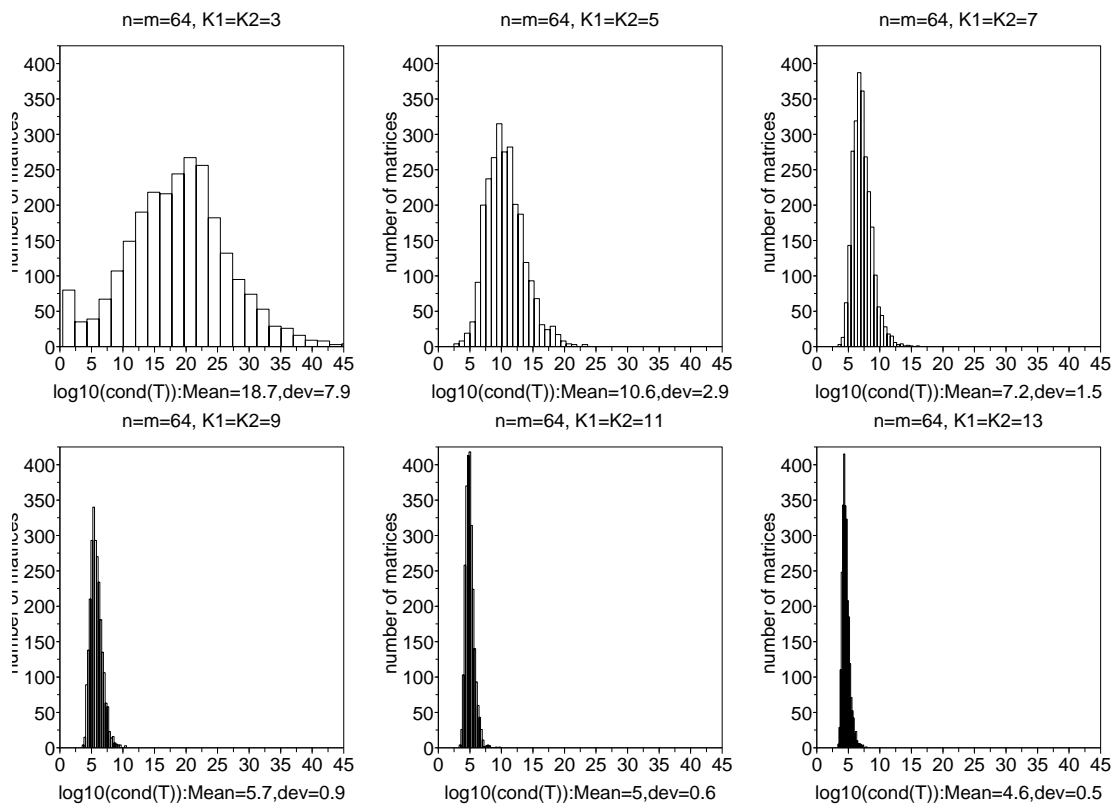


FIG. 3 – Les matrices ici sont de taille $64^2 \times 64^2$. La largeur des bandes et le nombre des essais sont comme pour Fig. 2

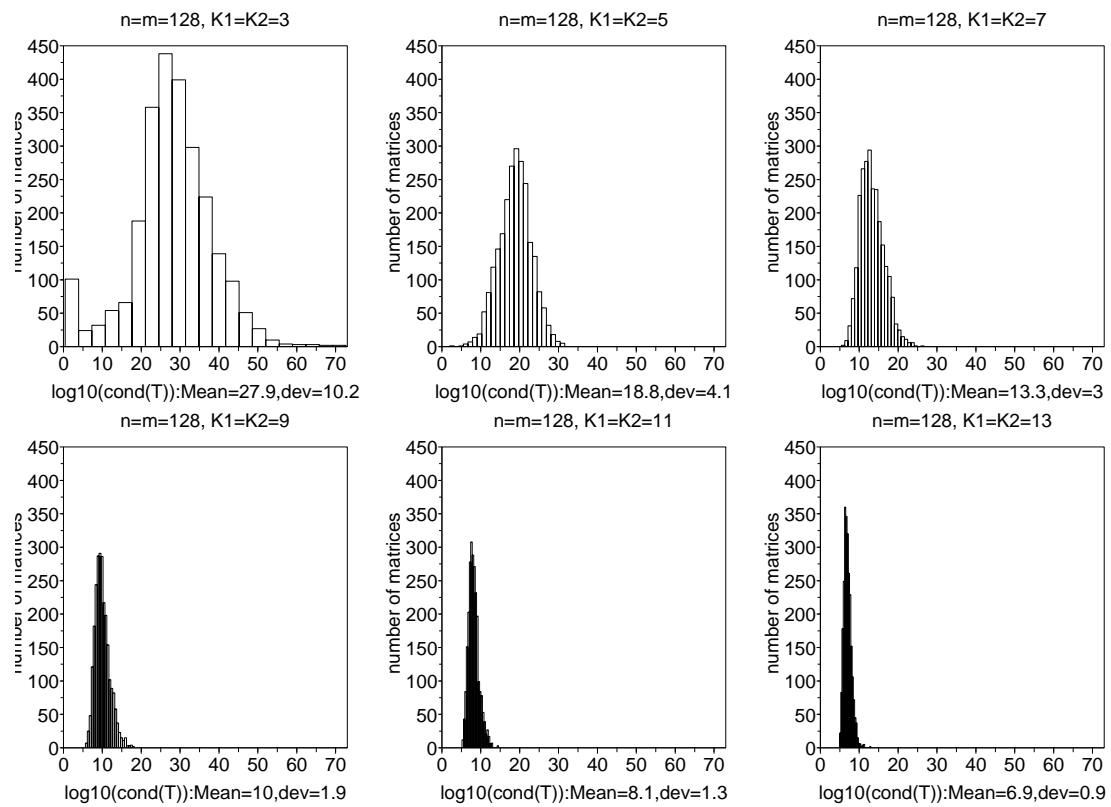


FIG. 4 – Matrices de Taille $128^2 \times 128^2$.

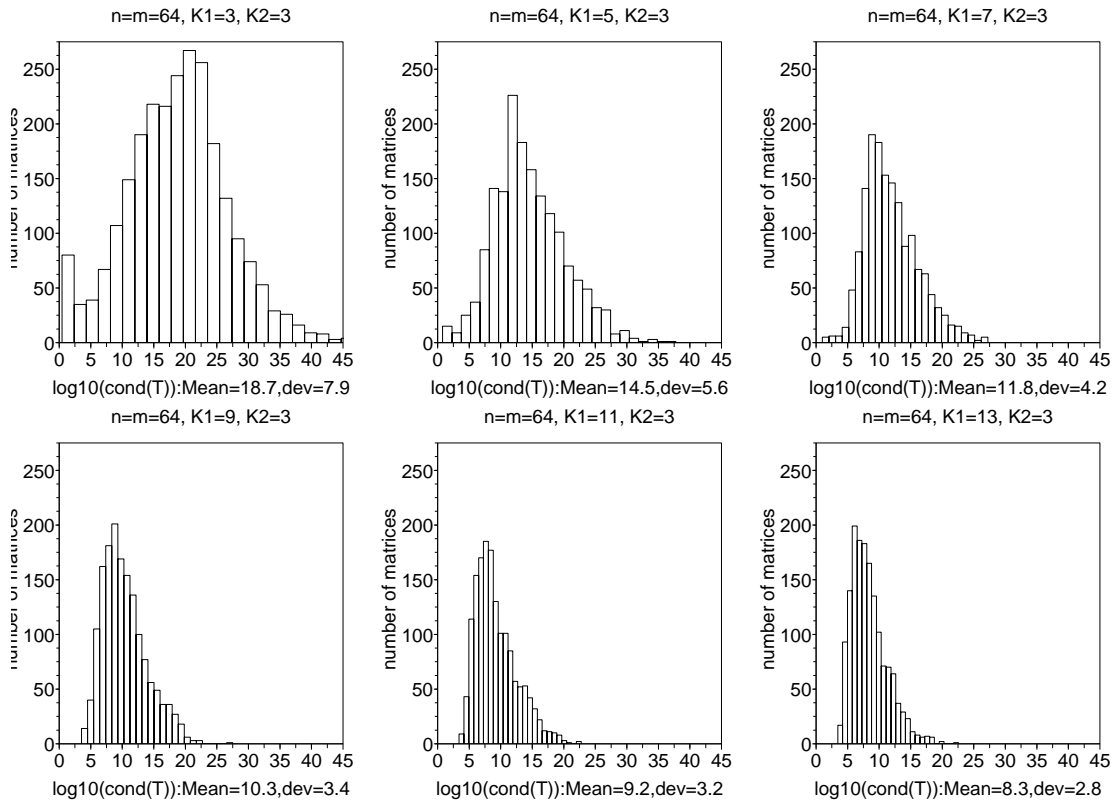


FIG. 5 – Les matrices sont de taille $64^2 \times 64^2$. Ici $k_1 = 1$ est fixe, $K_2 = 2k_2 + 1$ varie entre 3 et 13. Cet histogramme met en évidence la dépendance entre largeur des bandes et nombre de conditionnement quand la largeur d'une seule bande est variable.

Démonstration. La décomposition suivante est immédiate :

$$\begin{aligned}
A &= \begin{pmatrix} a_0 & a_{-1} & \dots & a_{-k} & 0 & \dots & 0 \\ a_1 & a_0 & \dots & a_{-k+1} & a_{-k} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_k & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-k} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-1} \\ 0 & \dots & 0 & a_k & \dots & a_1 & a_0 \end{pmatrix} \\
&= \begin{pmatrix} a_0 & \dots & a_{-k} & 0 & a_k & \dots & a_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_k & \ddots & \ddots & \ddots & \ddots & \ddots & a_k \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_{-k} & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-k} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{-1} & \dots & a_{-k} & 0 & a_k & \dots & a_0 \end{pmatrix} - \begin{pmatrix} 0 & \dots & 0 & a_k & \dots & a_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & a_k \\ a_{-k} & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{-1} & \dots & a_{-k} & 0 & \dots & 0 \end{pmatrix} \\
&= C((a_0 \dots a_k 0 \dots 0 a_{-k} \dots a_{-1})^T) + R = C + R.
\end{aligned} \tag{4}$$

□

Corollaire 1. *Supposons que A vérifie les hypothèses de la proposition 1. Alors, il existe une méthode directe pour résoudre le système $Ax = b$ en $\mathcal{O}(n \log n) + \mathcal{O}(nk \log k) + \mathcal{O}(k^3)$ opérations.*

Démonstration. En écrivant $R = GH^T$, avec $H, G \in \mathbb{K}^{n \times r}$ ($r = 2k$), et en utilisant la formule de Sherman-Morrison-Woodbury sur le système $(C+R)x = b$ on obtient

$$x = C^{-1}b - C^{-1}G(I_r + H^T C^{-1}G)^{-1}H^T C^{-1}b.$$

Remarquons tout d'abord que G est creuse avec $k(k+1)$ éléments non nuls :

$$G = \begin{pmatrix} 0_{k \times k} & g \\ 0_{(n-2k) \times k} & 0_{(n-2k) \times k} \\ f & 0_k \end{pmatrix} \text{ et } H = \begin{pmatrix} I_k & 0_{k \times k} \\ 0_{(n-2k) \times k} & 0_{(n-2k) \times k} \\ 0_{k \times k} & I_k \end{pmatrix},$$

avec $f = L(a_{-k} \dots a_{-1})^T$ et $g = U(a_k \dots a_1)^T$, et $L(v)$ (resp. $U(v)$) la matrice de Toeplitz triangulaire inférieure (resp. triangulaire supérieure) de première colonne (resp. première ligne) égale à v . Ainsi la multiplication d'une matrice de taille $n \times n$ par G coûte $\mathcal{O}(nk^2)$ opérations (la multiplication d'une matrice creuse qui contient p éléments non nuls par un vecteur coûte $2p$ opérations). On peut même faire mieux, en multipliant G par une matrice $n \times n$ en $\mathcal{O}(nk \log k)$ comme f et g sont de Toeplitz.

Par suite x est obtenu en faisant :

- $v_1 = C^{-1}b : \mathcal{O}(n \log n)$ opérations,
- $v_2 = H^T v_1 : 0$ opérations, parcequ'il n'y a pas que des 0 et des 1 dans H ,
- $C^{-1}v_1$, avec $v_1 = G(I_r + H^T C^{-1}G)^{-1} H^T C^{-1}b$.
- $H^T C^{-1}G : \mathcal{O}(n \log n) + \mathcal{O}(nk \log k)$ opérations (calculer C^{-1} en $\mathcal{O}(n \log n)$ opérations puis la multiplier par G en $\mathcal{O}(nk \log k)$ opérations),
- $v_3 = (I_r + H^T C^{-1}G)^{-1} v_2 : \mathcal{O}(k^3)$ opérations,
- $v_4 = G v_3 : \mathcal{O}(k \log k)$ opérations,
- $v_5 = C^{-1} v_4 : \mathcal{O}(n \log n)$ opérations.

□

Remarque 1. On a $\mathcal{O}(nk \log k)$ est plus 'grand' que $\mathcal{O}(n \log n)$ si $k = \mathcal{O}(n^\alpha)$ pour un $\alpha \in [0, 1[$.

3.1.1 Instabilité de l'algorithme

Dans leur livre, *Polynomial and matrix computation*, Victor Pan et Dario Bini, énoncent que cet algorithme rencontre des problèmes de stabilité. Les problèmes peuvent être dûs à l'instabilité propre du système linéaire. En effet, on pourra constater sur les figures que la comparaison entre nombre de conditionnement et erreur est normale. Pour se faire, nous avons généré des matrices de Toeplitz (et de Toeplitz bande par blocs Toeplitz bandes. Nous avons testé l'algorithme sur les deux types de matrices) bande pseudo aléatoires uniformes. Nous avons également généré des vecteurs x aléatoires et nous avons calculé l'erreur entre x et \tilde{x} qui est la *solution numérique*, via notre algorithme, de $T\tilde{x} = b$, b étant égal à Tx . Cette comparaison semble montrer que l'instabilité ne vient pas de l'algorithme.

3.2 Plongement dans une matrice engendrée par $Z + Z^T$

On va décrire maintenant un deuxième algorithme de résolution d'une matrice de Toeplitz bande. Dans cette section on travaille avec des matrices carrées de taille n . Soit τ_n (τ s'il y a pas de confusion) l'algèbre engendrée par $W = Z + Z^T$. Les matrices de Toeplitz dans cette section sont symétriques.

Proposition 2. Soit $A \in \tau$, alors ses coefficients vérifient

$$\begin{aligned} a_{i-1,j} + a_{i+1,j} &= a_{i,j+1} + a_{i,j-1} \\ a_{0,j} &= a_{n+1,j} = a_{i,0} = a_{i,n+1} = 0 \end{aligned} \quad (5)$$

Démonstration. Si A est dans τ , elle est de la forme $A = \sum_{i=0}^{n-1} \alpha_i W^i$. Or $W^{k+1} = W^k \cdot W = W \cdot W^k$ et si M une matrice alors $(MW)_{ij} = M_{i,j+1} + M_{i,j-1}$ et $(WM)_{i,j} = M_{i-1,j} + M_{i,j+1}$. Donc

$$(W^{k+1})_{ij} = (W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j}. \quad (6)$$

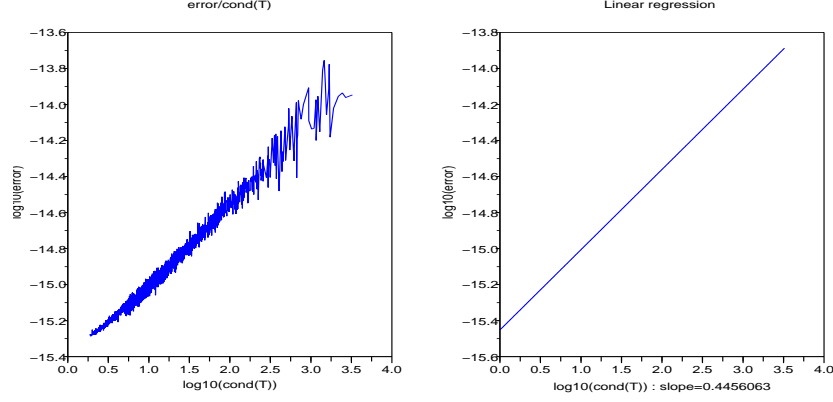


FIG. 6 – Les matrices ici sont de Toeplitz bande de taille 400×400 et la largeur de la bande $2k + 1 = 5$. Pour 3000 essais, on trace le logarithme à base 10 de l'erreur dû à notre algorithme par rapport au logarithme à base 10 du nombre de conditionnement. On trouve que la pente de la droite de régression est plus petit que 0.5 !

Montrons par récurrence sur k que chaque W^k vérifie la propriété (5) : W la vérifie. Supposons que W^k est telle que

$$(W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j},$$

D'après (6) :

$$(W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j-1} + (W^k)_{i+1,j-1} + (W^k)_{i-1,j+1} + (W^k)_{i+1,j+1},$$

et

$$(W^k)_{i-1,j} + (W^k)_{i+1,j} = (W^k)_{i-1,j-1} + (W^k)_{i-1,j+1} + (W^k)_{i+1,j-1} + (W^k)_{i+1,j+1},$$

$$\text{par suite } (W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j}. \quad \square$$

Corollaire 2. De là nous déduisons que l'algèbre τ_n peut être identifiée à

$$B = \{B_k \in \tau; B_k e_1 = e_k, 1 \leq k \leq n\}.$$

Pour $A \in \tau$,

$$A = \sum_{k=1}^n a_k B_k,$$

avec $(a_1, \dots, a_n)^T$ la première ligne de A .

Démonstration. Pour construire une matrice de l'algèbre τ il suffit de connaître sa première colonne. Or, comme B_k , $1 \leq k \leq n$, est la matrice de τ_n telle que sa première colonne est e_k , et comme la première colonne de A vaut $\sum_{k=1}^n a_k e_k$, alors $A = \sum_{k=1}^n a_k B_k$. \square

Proposition 3. *Soit Δ l'espace vectoriel des matrices de Toeplitz $n \times n$, de largeur de bande $2k + 1$. Soit dans $\tau_{n+2(k+1)}$, le sous espace vectoriel engendré par les matrices B_i telles que $B_i e_1 = e_i - e_{i-2}$, avec $i = 3 \dots k + 1$, ainsi que $B_1 = I, B_2 = W$. Soit $E = \{k + 1, \dots, n - k - 1\}$. Alors les matrices $\tilde{B}_i = ((B_i)_{lp})_{l,p \in E}$ forment une base de Δ .*

Démonstration. En utilisant la technique donnée dans la proposition précédente pour calculer un élément dans τ , on remarque que B_i est de la forme :

$$\begin{array}{cccccc|cccc|c}
 0 & \dots & -1 & 0 & 1 & 0 & & \dots & & 0 & \\
 \vdots & & -1 & & & 1 & & & & & \\
 -1 & & & & & & 1 & & & & \bigcirc \\
 0 & & & & & & & \ddots & & & \\
 1 & & & & & & & & 1 & & \\
 \hline
 0 & 1 & & & & & & & & 1 & \\
 & & \ddots & & & & & & & \ddots & \\
 & & & 1 & & & & & & & 1 \\
 & \vdots & & & 1 & & & & & & 1 \\
 & & & & & 1 & & & & \ddots & \\
 & & & & & \ddots & & & & & 1 \\
 & & & & & & 1 & & & & \\
 \hline
 & \ddots & & & \ddots & & \ddots & \ddots & \ddots & & \ddots
 \end{array}$$

FIG. 7 – La matrice B_i comporte un centre, c'est à dire l'intersection des lignes et des colonnes indexées par $E = \{k + 1, \dots, n - k - 1\}$, et une périphérie qui est son complémentaire

et \tilde{B}_i est la matrice qui contient des 1 sur les diagonales qui commencent par les éléments $(1, i)$ et $(i, 1)$ respectivement. Par suite les \tilde{B}_i génèrent Δ \square

Soit $T \in \Delta$ telle que $T = \sum_{i=1}^{k+1} a_i \tilde{B}_i$. On peut donc la plonger dans une matrice $M \in \tau$ avec $M = \sum_{i=1}^{k+1} a_i B_i$ de taille $n + 2(k + 1)$. M a donc la forme suivante :

$$M = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12}^T & T & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{pmatrix},$$

avec M_{11}, M_{13} et M_{33} de taille $(k+1) \times (k+1)$, M_{12} et M_{23}^T de taille $(k+1) \times n$. On cherche à résoudre le système $Tx = b$. Étudions le système suivant :

$$\begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12}^T & T & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{pmatrix} \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix} = \begin{pmatrix} b_1 \\ b \\ b_3 \end{pmatrix}.$$

Dans ce système, les inconnues sont x , b_1 et b_3 et on a $Tx = b$. Pour trouver b_1 et b_3 on procède comme suit : soit

$$M^{-1} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12}^T & \mu_{22} & \mu_{23} \\ \mu_{13}^T & \mu_{23}^T & \mu_{33} \end{pmatrix}.$$

Les vecteurs b_1 et b_3 vérifient :

$$\begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12}^T & \mu_{22} & \mu_{23} \\ \mu_{13}^T & \mu_{23}^T & \mu_{33} \end{pmatrix} \begin{pmatrix} b_1 \\ b \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix},$$

et par suite résolvent le système

$$\begin{cases} \mu_{11}b_1 + \mu_{13}b_3 = -\mu_{12}b, \\ \mu_{13}^Tb_1 + \mu_{33}b_3 = -\mu_{23}^Tb. \end{cases}$$

Ce système, de taille $2(k+1) \times 2(k+1)$, donne b_1 et b_3 en $\mathcal{O}(k^3)$ opérations. Pour construire ce système on a besoin aussi de calculer $\mu_{11}, \mu_{12}, \mu_{13}$ et μ_{23} , puis de calculer $\mu_{12}b$ et μ_{23}^Tb . comme μ_{12} et μ_{23}^T sont de taille $(k+1) \times n$, ce calcul va coûter $\mathcal{O}(nk)$ opérations. Le calcul de $\mu_{11}, \mu_{12}, \mu_{13}, \mu_{23}$ coûte $\mathcal{O}(n \log^2(n)) + \mathcal{O}(k^2n)$ opérations (voir corollaire ci-dessus).

Proposition 4. Soit $M \in \tau_n$. On peut résoudre le système $Mx = b$ en $\mathcal{O}(n \log^2 n)$ opérations.

Démonstration. $M \in \tau$ Comme M est dans τ_n , elle est de la forme

$$M = \sum_{i=1}^{n-1} m_i W^i.$$

Or les valeurs propres et vecteurs propres de W sont donnés par :

$$\lambda_k = 2 \cos \frac{k\pi}{n+1}, \quad v_k = \left(\sin \frac{jk\pi}{n+1} \right)_{1 \leq j \leq n}, \quad k = 1 \dots n,$$

comme on peut le vérifier simplement. Soit donc s la matrice de la transformation de Fourier en sinus :

$$s = \begin{pmatrix} \sin \frac{\pi}{n+1} & \dots & \sin \frac{n\pi}{n+1} \\ \vdots & \ddots & \vdots \\ \sin \frac{n\pi}{n+1} & \dots & \sin \frac{n^2\pi}{n+1} \end{pmatrix} = \left(\sin \frac{ij\pi}{n+1} \right)_{1 \leq i, j \leq n}.$$

Classiquement, $s^{-1} = 2s/(n+1)$. Si on pose $S = s\sqrt{2/(n+1)}$, alors $S^{-1} = S$ et $W = SDS$ avec D la matrice diagonale des valeurs propres. Nous en déduisons la diagonalisation de M :

$$M = S \left(\sum_{i=0}^{n-1} m_i D^i \right) S.$$

Le calcul de la somme des $m_i D^i$ se fait en $\mathcal{O}(n \log^2 n)$ (évaluation d'un polynôme (avec coefficients m_i) aux n valeurs propres λ_k). \square

Corollaire 3. *Soit $M \in \tau_{n+2(k+1)}$. Pour calculer la périphérie de taille $k+1$ de M^{-1} on a besoin de $\mathcal{O}((n+k) \log^2(n+k)) + \mathcal{O}(nk^2)$ opérations.*

Démonstration. Notons $K = k+1$. La matrice M est symétrique, elle est aussi symétrique par rapport à l'antidiagonale. Donc, pour calculer les K premières lignes, les K premières colonnes, les K dernières lignes et les K dernières colonnes de M^{-1} (la périphérie de M^{-1}) il suffira de calculer les K premières colonnes. On a donc besoin de résoudre K systèmes linéaires avec la matrice M . D'après la proposition précédente ce calcul coûtera $\mathcal{O}(K(n+2K) \log^2(n+2K))$ opérations.

On peut construire la périphérie d'une autre manière. En effet, comme $M^{-1} \in \tau$, on peut calculer sa première colonne en $\mathcal{O}((n+2K) \log(n+2K))$ opérations, et construire les autres colonnes demandées en utilisant la technique de la proposition (2), ce qui demande $\mathcal{O}(nk^2)$ opérations. \square

Corollaire 4. *La résolution de $Tx = b$ coûte $\mathcal{O}(n \log^2 n) + \mathcal{O}(k^3)$ opérations.*

3.3 Plogement dans une matrice circulante

L'idée de cet algorithme est de plonger la matrice de Toeplitz bande dans une matrice circulante à la place de la plonger dans une matrice de l'algèbre τ . Dans cette section, les matrices de toeplitz bandes ne sont pas nécessairement symétriques.

Proposition 5. *Soit $A \in \mathbb{K}^{n \times n}$ une matrice de Toeplitz bande, de largeur de bande $2k+1$. Alors on peut plonger A dans une matrice circulante de taille $(n+k) \times (n+k)$.*

Proposition 6. *Soit A une matrice de Toeplitz bande donnée comme dans (4). On peut donc plonger A dans la matrice circulante $C = C(r)$ de première colonne r donné par*

$$r = (\underbrace{a_0, \dots, a_k, 0, \dots, 0}_n, a_{-k}, \dots, a_{-1})^T.$$

En effet,

$$\begin{aligned}
C &= \left(\begin{array}{cccccc|cccc}
a_0 & \dots & a_{-k+1} & a_{-k} & 0 & \dots & 0 & a_k & \dots & a_1 \\
\vdots & \ddots & & & \ddots & & & & \ddots & \vdots \\
a_{k-1} & & \ddots & & & \ddots & & & & a_k \\
a_k & & & \ddots & & & a_{-k} & & & \\
0 & \ddots & & & \ddots & & & a_{-k} & & \\
\vdots & & \ddots & & & \ddots & \vdots & \vdots & \ddots & \\
0 & & & a_k & \dots & a_0 & a_{-1} & \dots & a_{-k} \\
\hline
a_{-k} & & & & a_k & \dots & a_1 & a_0 & \dots & a_{-k+1} \\
\vdots & \ddots & & & & \ddots & \vdots & \vdots & \ddots & \vdots \\
a_{-1} & \dots & a_{-k} & & & & a_k & a_{k-1} & \dots & a_0
\end{array} \right) \quad (7) \\
&= \left(\begin{array}{c|c}
A & B \\
\hline
D & T
\end{array} \right)
\end{aligned}$$

avec T une matrice de Toeplitz de taille $k \times k$, D et B^T sont de Toeplitz, creuses, de taille $k \times n$.

Comme dans l'algorithme précédent, on réduit la résolution du système $Ax = b$ à la résolution du système

$$C \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

avec c un vecteur inconnu de longueur k .

En écrivant

$$C^{-1} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix},$$

avec γ_{11} de taille $n \times n$ et γ_{22} de taille $k \times k$, alors le vecteur c résolve le système de Toeplitz, de taille $k \times k$, suivant :

$$\gamma_{22}c = -\gamma_{21}b.$$

On peut donc calculer c en $\mathcal{O}(k \log^2 k)$ flops. Le vecteur $(x \ 0)^T$ est la solution d'un système circulant. Donc, une fois calculer c , $(x \ 0)^T$ sera calculé en $\mathcal{O}((n+k) \log(n+k))$ flops. Cet algorithme est un peu plus rapide que les deux premiers parceque le calcul de c peut se faire en $\mathcal{O}(k \log^2 k)$ flops.

Proposition 7. La résolution du système $Ax = b$ coûte $\mathcal{O}(n \log n) + \mathcal{O}(kn) + \mathcal{O}(k \log^2 k)$ flops.

Démonstration. Le calcul de $\gamma_{21}b$ demande $\mathcal{O}(kn)$ flops. Puis le calcul de c demande $\mathcal{O}(k \log^2 k)$, et à la fin la résolution du système $C \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$ se fait en $\mathcal{O}(n \log n)$ flops. \square

4 Cas par blocs

Le cas par blocs est une généralisation directe du cas scalaire mais où y rencontre quelques complications, et les comptes d'opérations sont différents.

4.1 Transformation en matrice circulante plus matrice de petit rang

Soit T une matrice comme dans (1) et (2).

Proposition 8. *On peut décomposer T en $T = C + R$, avec C une matrice circulante par blocs circulants et R de rang au plus $2(k_1n + k_2m)$, et au plus $\mathcal{O}(k_1^2k_2n + k_2^2k_1m)$ éléments non nuls.*

Démonstration. Écrivons $T_i = C_i + \tilde{R}_i$, avec $-k_1 \leq i \leq k_1$, où C_i est une matrice circulante et \tilde{R}_i une matrice de rang au plus $2k_2$. Nous pouvons alors décomposer T comme suit :

$$\begin{aligned}
 T &= \begin{pmatrix} C_0 & \dots & C_{-k_1} & 0 & C_{k_1} & \dots & C_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & C_{k_1} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ C_{-k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & C_{-k_1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{-1} & \dots & C_{-k_1} & 0 & C_{k_1} & \dots & C_0 \end{pmatrix} - \\
 &\quad \begin{pmatrix} 0 & \dots & 0 & C_{k_1} & \dots & C_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & C_{k_1} \\ C_{-k_1} & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{-1} & \dots & C_{-k_1} & 0 & \dots & 0 \end{pmatrix} - \begin{pmatrix} \tilde{R}_0 & \dots & \tilde{R}_{-k_1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \tilde{R}_{k_1} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \tilde{R}_{-k_1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \tilde{R}_{k_1} & \dots & \tilde{R}_0 \end{pmatrix} \\
 &= C + R_1 + R_2 = C + R.
 \end{aligned}$$

Nous remarquons que R_1 est au plus de rang $2k_1n$ et R_2 est de rang $2k_2m$.

Dans R_1 il y a $2 \times k_1(k_1 + 1)/2$ blocs non nuls, et dans chaque bloc C_i il y a $n + 2(n-1) + 2(n-2) + \dots + 2(n-k_2) + (k_2^2 + k_2) = 2k_2n + n$ éléments non nuls. Ainsi R_1 contient $(k_1^2 + k_1)(2k_2n + n) = \mathcal{O}(k_1^2k_2n)$ éléments non nuls ; la situation est analogue pour R_2 . \square

En écrivant $R = GH^T$, avec $G, H \in \mathbb{K}^{N \times r}$ ($r = 2(k_1n + k_2m)$), et en appliquant la formule de Sherman-Morrison-Woodbury sur le système $(C +$

$R)x = b$ on obtient

$$x = C^{-1}b - C^{-1}G(I_r + H^T C^{-1}G)^{-1}H^T C^{-1}b.$$

Proposition 9. *G est creuse avec $\mathcal{O}(k_1^2 k_2 n) + \mathcal{O}(k_2^2 k_1 m)$ éléments non nuls, et H ne comprend que des 0 et des 1.*

Démonstration. Les matrices G et H se décomposent en

$$G = (G_1 \ G_2) \text{ et } H^T = \begin{pmatrix} H_1^T \\ H_2^T \end{pmatrix}$$

avec

$$G_1 = \begin{pmatrix} 0_{nk_1} & E \\ \vdots & 0 \\ 0 & \vdots \\ F & 0_{nk_1} \end{pmatrix},$$

pour décrire la matrice G_2 , notons

$$E_l = \{(l-1)n+1, \dots, (l-1)n+k_2\} \cup \{ln-k_2+1, \dots, ln\}$$

et

$$E = \cup_{l=1}^m E_l$$

alors

$$G_2 = ((R_2)_{ij})_{\substack{1 \leq i \leq mn \\ j \in E}},$$

$$H_1^T = \begin{pmatrix} I_{nk_1} & 0 & \dots & 0_{nk_1} \\ 0_{nk_1} & \dots & 0 & I_{nk_1} \end{pmatrix},$$

et

$$H_2^T = \left(\begin{array}{cc|cc|cc} I_{k_2} & 0_{k_2 \times (n-k_2)} & \dots & \dots & I_{k_2} & 0_{k_2 \times (n-k_2)} \\ 0_{k_2 \times (n-k_2)} & I_{k_2} & \dots & \dots & 0_{k_2 \times (n-k_2)} & I_{k_2} \end{array} \right)$$

chaque bloc dans H_2 est de taille $2k_2 \times m$.

Donc, en utilisant le compte fait dans la proposition précédente on voit que G contient $(k_1^2 + k_1)(2k_2n + n) + (k_2^2 + k_2)(2k_1m + m) = \mathcal{O}(k_1^2 k_2 n) + \mathcal{O}(k_2^2 k_1 m)$. \square

Corollaire 5. *La multiplication de G par un vecteur coûte $\mathcal{O}(k_1^2 k_2 n) + \mathcal{O}(k_2^2 k_1 m)$ opérations.*

Corollaire 6. *En supposant que k_1 et k_2 sont petit devant m et n respectivement alors la résolution du système $Tx = b$ donné en (1) et (2) coûte $\mathcal{O}(N^{3/2})$ opérations.*

Démonstration. Soit $N = nm$, $K = k_1^2 k_2 n + k_2^2 k_1 m$, et $r = 2(k_1 n + k_2 m)$.

On a $x = C^{-1}b - C^{-1}G(I_r + H^T C^{-1}G)^{-1}H^T C^{-1}b$, donc x est obtenu en faisant :

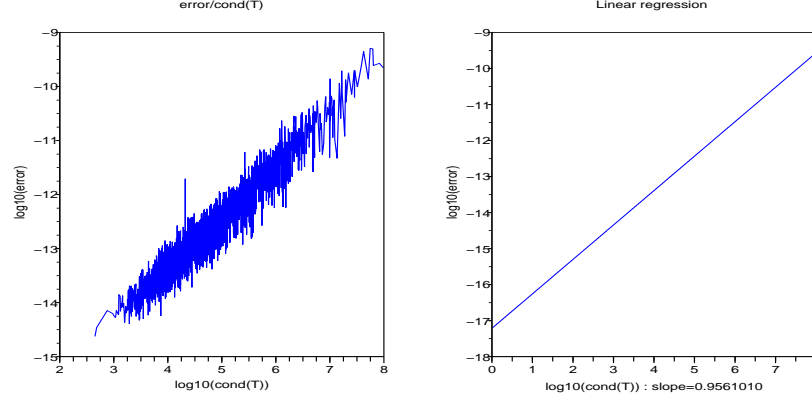


FIG. 8 – Les matrices sont de Toeplitz bande par blocs Toeplitz bande de taille $32^2 \times 32^2$ de largeur de bande $K_1 = 2k_1 + 1 = K_2 = 2k_2 + 1 = 5$. Pour 3000 essais, on trace le logarithme à base 10 de l'erreur dû à notre algorithme par rapport au logarithme à base 10 du nombre de conditionnement. On trouve que la pente de la droite de régression est à peu près égale 1.

- $v_1 = C^{-1}b : \mathcal{O}(N \log N)$.
- $v_2 = H^T v_1 : 0$ opération, car il n'y a que des 0 et des 1 dans H .
- $H^T C^{-1}G : \mathcal{O}(N \log N) + \mathcal{O}(NK) = \mathcal{O}(N \log N) + \mathcal{O}(N^{3/2})$, le calcul de C^{-1} demande $\mathcal{O}(N \log N)$ opérations et la multiplication par G demande $\mathcal{O}(NK)$ opérations car G est creuse.
- $v_3 = (I_r + H^T C^{-1}G)^{-1}v_2 : \mathcal{O}(r^3) = \mathcal{O}(N^{3/2})$, résolution classique.
- $v_4 = Gv_3 : \mathcal{O}(K) = \mathcal{O}(N^{1/2})$, multiplication de G par un vecteur, donc proportionnel au nombre d'éléments non nuls dans G .
- $v_5 = C^{-1}v_4 : \mathcal{O}(N \log N)$.

□

4.1.1 Stabilité

Comme dans le cas scalaire, on a appliqué cet algorithme sur des matrices de Toeplitz bande par blocs Toeplitz bandes pseudo aléatoires pour obtenir les erreurs entre la solution de $Tx = b$ calculée par notre algorithme et la solution exacte. En traçant ces erreurs par rapport aux nombres de conditionnement, on a obtenu une ligne de régression de pente $\simeq 1$. Ce qui prouve qu'il y a pas un problème de stabilité pour cet algorithme. Mais on remarque que la pente est significativement augmenté de $\simeq 0.5$ pour le cas scalaire à $\simeq 1$ pour le cas par blocs!!!

4.2 Plongement dans une matrice engendrée par $Z + Z^T$

Dans cette section T est une matrice de Toeplitz, par blocs de Toeplitz symétrique par blocs, et les blocs sont symétriques, bande par blocs, et les blocs sont aussi bandes. Les entiers m, n, k_1 et k_2 sont les dimensions utilisées dans les sections précédents. Pour résoudre le système $Tx = b$, on va essayer de plonger la matrice T dans une matrice de l'algèbre $\tau_{m,n}$ (ou τ s'il y a pas de confusion) engendrée par $W = (Z_m + Z_m^T) \otimes (Z_n + Z_n^T)$. Une matrice $M \in \tau$ est donc de la forme

$$M = \sum_{\substack{0 \leq i \leq m-1 \\ 0 \leq j \leq n-1}} \alpha_{ij} (Z_m + Z_m^T)^i \otimes (Z_n + Z_n^T)^j = \sum \alpha_{i,j} W^{(i,j)}.$$

Si M est de plus bande par blocs bandes, les bornes supérieures respectives de i et j dans la sommation sont k_1 et k_2 . On va essayer de procéder comme dans le cas scalaire pour avoir un algorithme rapide.

Proposition 10. *Soit M dans l'algèbre τ . On peut résoudre le système $Mx = b$ en $\mathcal{O}(nm \log^2 mn)$ opérations.*

Démonstration. Notons s_k la matrice de la transformation en sinus de dimension k , $S_k = s_k \sqrt{2/(k+1)}$ et D_k la matrice diagonale $\text{diag}(\lambda_1, \dots, \lambda_k)$. Comme $M \in \tau$ alors $M = \sum \alpha_{ij} W^{(i,j)}$ avec $W := W_m \otimes W_n$ et $W^{(i,j)} := (Z_m + Z_m^T)^i \otimes (Z_n + Z_n^T)^j$. Or $W = W_m \otimes W_n = (S_m D_m S_m) \otimes (S_n D_n S_n) = (S_m \otimes S_n)(D_m \otimes D_n)(S_m \otimes S_n) =: SDS$. Donc M admet la diagonalisation :

$$M = SES \text{ avec } E = \sum_{\substack{0 \leq i \leq m-1 \\ 0 \leq j \leq n-1}} \alpha_{ij} D_m^i \otimes D_n^j.$$

E peut être calculé en $\mathcal{O}(mn \log^2 mn)$ opérations : il s'agit d'évaluer un polynôme à deux variable en mn points. Comme M^{-1} admet la décomposition $M^{-1} = SE^{-1}S$, la multiplication de M^{-1} par un vecteur coûte 2 transformées rapide en sinus et une multiplication par une matrice diagonale, le tout en $\mathcal{O}(mn \log mn)$ opérations. \square

Soit T une matrice bande Toeplitz symétrique par blocs bande Toeplitz symétrique, comme en (1). De même technique que pour le cas scalaire, nous la plongeons dans une matrice $M \in \tau$, en la plongeant tout d'abord par blocs, puis par blocs et après en plongeant chaque bloc.

La matrice M obtenue est une matrice de taille $m + 2(k_1 + 1)$ par blocs, et chaque bloc est de taille $n + 2(k_2 + 1)$; donc M est de taille $(m + 2(k_1 + 1))(n + 2(k_2 + 1))$.

$2(k_2 + 1) \simeq mn + 2mk_1 + 2nk_2 + 4k_1k_2$. Elle a la forme suivante :

$$M = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & \begin{array}{|c|c|c|} \hline \times & \times & \times \\ \hline \times & T_0 & \times \\ \hline \times & \times & \times \\ \hline \vdots & & \\ \hline \times & \times & \times \\ \hline \times & T_{n-1} & \times \\ \hline \times & \times & \times \\ \hline \end{array} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix}, \quad (8)$$

avec $M_{11}, M_{13}, M_{31}, M_{33}$ de taille $(k_1 + 1)(n + 2(k_2 + 1)) \times (k_1 + 1)(n + 2(k_2 + 1))$ et M_{12}, M_{32} sont de taille $(k_1 + 1)(n + 2(k_2 + 1)) \times m(n + 2(k_2 + 1))$.

On cherche à résoudre $Tx = b$ avec

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \text{les } x_i \text{ et les } b_i \text{ étant des vecteurs de taille } n.$$

Comme dans le cas scalaire, on va compléter le vecteur x par de nouvelles composantes nulles, dont les numéros de ligne sont les numéros des nouvelles ligne de M par rapport à T . Il faudra donc ajouter des lignes à b , ce qui créera de nouvelles inconnues :

$$\tilde{x} = \begin{pmatrix} 0_k \\ \bar{x} \\ 0_k \end{pmatrix} \quad \text{et} \quad \tilde{b} = \begin{pmatrix} \hat{b}_1 \\ \bar{b} \\ \hat{b}_2 \end{pmatrix},$$

avec $k = (k_1 + 1)(n + (k_2 + 1))$ et

$$\bar{x} = \begin{pmatrix} 0_{\bar{k}_2} \\ x_1 \\ \hline 0_{\bar{k}_2} \\ \vdots \\ \hline 0_{\bar{k}_2} \\ x_m \\ \hline 0_{\bar{k}_2} \end{pmatrix} \quad \text{et} \quad \bar{b} = \begin{pmatrix} \hat{b}_{11} \\ b_1 \\ \hline \hat{b}_{13} \\ \vdots \\ \hline \hat{b}_{m1} \\ b_m \\ \hline \hat{b}_{m3} \end{pmatrix},$$

avec $\bar{k}_2 = (k_2 + 1)$, les \hat{b}_{i1} et les \hat{b}_{i3} de taille \bar{k}_2 . Soit

$$\hat{b} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_{11} \\ b_1 \\ \hat{b}_{13} \\ \vdots \\ \hat{b}_{m1} \\ b_m \\ \hat{b}_{m3} \\ \hat{b}_3 \end{pmatrix}$$

Comme dans le cas scalaire, écrivons $\tilde{x} = M^{-1}\tilde{b}$, et en ne gardant dans ce système que les lignes nulles dans \tilde{x} et que les nouvelles inconnues, nous obtenons un système par rapport à \hat{b} , dont la matrice est de taille $\simeq nk_1 + mk_2$. Pour décrire ce système on va écrire M^{-1} d'une façon équivalente à l'écriture de M donnée en (8), elle sera donc donnée comme suit :

$$M^{-1} = \left(\begin{array}{c|c|c} \mu_{11} & \mu_{12} & \mu_{13} \\ \hline \mu_{21} & \left(\begin{array}{c|c|c} \mu_{11}^{ij} & \mu_{12}^{ij} & \mu_{13}^{ij} \\ \hline \mu_{21}^{ij} & \mu_{22}^{ij} & \mu_{23}^{ij} \\ \hline \mu_{31}^{ij} & \mu_{32}^{ij} & \mu_{33}^{ij} \end{array} \right)_{i,j=1\dots m} & \mu_{23} \\ \hline \mu_{31} & \mu_{32} & \mu_{33} \end{array} \right),$$

on découpe aussi μ_{12} (pareil pour μ_{32}) de la façon suivante :

$$\mu_{12} = (\mu_{12,11} \mid \mu_{12,1} \mid \mu_{12,13} \mid \dots \mid \mu_{12,m1} \mid \mu_{12,m} \mid \mu_{12,m3}).$$

Le système à résoudre sera donné par :

$$\left\{ \begin{array}{l} \mu_{11}\hat{b}_1 + \sum_{i=1}^m (\mu_{12,i1}\hat{b}_{i1} + \mu_{12,i3}\hat{b}_{i3}) + \mu_{13}\hat{b}_3 = - \sum_{i=1}^m \mu_{12,i}b_i \\ \mu_{12,j1}^T\hat{b}_1 + \sum_{i=1}^m (\mu_{11}^{ij}\hat{b}_{i1} + \mu_{13}^{ij}\hat{b}_{i3}) + \mu_{23,j1}\hat{b}_3 = - \sum_{i=1}^m \mu_{12}^{ij}b_i \quad j = 1 \dots m \\ \mu_{12,j3}^T\hat{b}_1 + \sum_{i=1}^m (\mu_{31}^{ij}\hat{b}_{i1} + \mu_{33}^{ij}\hat{b}_{i3}) + \mu_{23,j3}\hat{b}_3 = - \sum_{i=1}^m \mu_{32}^{ij}b_i \quad j = 1 \dots m \\ \mu_{31}\hat{b}_1 + \sum_{i=1}^m (\mu_{32,i1}\hat{b}_{i1} + \mu_{32,i3}\hat{b}_{i3}) + \mu_{33}\hat{b}_3 = - \sum_{i=1}^m \mu_{12,i}b_i \end{array} \right.$$

Pour former le deuxième membre de ce système il faut $\mathcal{O}(k_1(n + k_2).m(n + 2k_2)) + \mathcal{O}(2m^2.k_2.n) \simeq \mathcal{O}(N^{3/2}) + \mathcal{O}(N^{3/2})$ opérations : 2 multiplications d'une matrice de taille $(k_1 + 1)(n + (k_2 + 1)) \times m(n + 2(k_2 + 1))$ par un vecteur et $2m^2$ multiplications matrice-vecteur avec des matrices de taille $(k_2 + 1)n$.

Proposition 11. *En supposant que k_1 et k_2 sont petit devant m et n respectivement alors le calcul de la périphérie (en deux dimension) de M^{-1} coûte $\mathcal{O}((n + k_1)(m + k_2) \log((n + k_1)(m + k_2))) + \mathcal{O}(N^{3/2})$ opérations.*

Démonstration. Même démonstration que pour le cas scalaire \square

Finalement, la résolution du système initiale requiert $\mathcal{O}(N^{3/2})$ opérations pour être résolu.

4.3 Plogement dans une matrice circulante par blocs circulants

Soit T une matrice bande Toeplitz par blocs bande Toeplitz comme en (1). On s'intéresse au problème $Tx = b$ avec

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \text{les } x_i \text{ et les } b_i \text{ étant des vecteurs de taille } n.$$

En suivant les techniques du cas scalaire, on peut plonger T dans une matrice circulante par blocs circulants, C , de taille $(m + k_1)(n + k_2) \times (m + k_1)(n + k_2)$. En effet, on plonge tout d'abord chaque bloc T_i , pour $-k_1 \leq i \leq k_1$, dans une matrice circulante de taille $(n + k_2) \times (n + k_2)$. Par suite on plonge par bloc la matrice obtenue, qui est une matrice bande Toeplitz par blocs, dans une matrice circulante par blocs. La matrice finale est bien une matrice circulante par blocs circulants de taille $(m + k_1)(n + k_2) \times (m + k_1)(n + k_2)$.

On suit les techniques de la section précédente : on pose le problème

$$C\tilde{x} = \tilde{b}$$

avec

$$\tilde{x} = \begin{pmatrix} \bar{x} \\ 0_k \end{pmatrix} \quad \text{et} \quad \tilde{b} = \begin{pmatrix} \bar{b} \\ c \end{pmatrix}$$

avec $k = k_1(n + k_2)$, c un vecteur inconnu de longueur k et

$$\bar{x} = \begin{pmatrix} x_1 \\ 0_{k_2} \\ \vdots \\ x_m \\ 0_{k_2} \end{pmatrix} \quad \text{et} \quad \bar{b} = \begin{pmatrix} b_1 \\ c_1 \\ \vdots \\ b_m \\ c_m \end{pmatrix}$$

où chaque c_i , $1 \leq i \leq m$, est un vecteur inconnu de longueur k_2 .

Comme dans le cas scalaire, pour trouver le vecteur inconnu

$$\zeta = \begin{pmatrix} c_1 \\ \vdots \\ c_m \\ c \end{pmatrix}$$

on inverse a matrice C et on extraire un système linéaire, $M\zeta = g$, dont la solution est ζ . Ce système sera de taille $K \times K$, avec $K = mk_2 + nk_1 + k_1k_2$. On peut remarquer quelque structure dans M , mais elle n'est pas TBT (dans le cas scalaire, on a obtenu un système de Toeplitz). Le calcul de g demande une multiplication d'une matrice de taille $k_1(n+k_2) \times m(n+k_2)$ par un vecteur, et m^2 multiplications des matrices de taille $k_2 \times n$ par des vecteurs. Ceci demande en total, pour calculer g , $\mathcal{O}(k_1mn^2 + 2k_1k_2mn + k_1k_2^2m + k_2m^2n)$ flops, qui est équivalent à $\mathcal{O}(N^{3/2})$ flops. La résolution du système $M\zeta = g$ demande $\mathcal{O}(K^3)$ flops, ce qui est aussi équivalent à $\mathcal{O}(N^{3/2})$ flops.

A la fin, la résolution du système $C\tilde{x} = \tilde{b}$ se fait en $\mathcal{O}((m+k_1)(n+k_2) \log(m+k_1)(n+k_2))$ flops, qui est équivalent à $\mathcal{O}(N \log N)$.

Cet algorithme est donc d'ordre $N^{3/2}$ flops.

Références

- [1] Greg W. Anderson and Ofer Zeitouni. A CLT for a band matrix model. *Probab. Theory Related Fields*, 134(2) :283–338, 2006.
- [2] Dario Bini and Victor Y. Pan. *Polynomial and matrix computations. Vol. 1.* Progress in Theoretical Computer Science. Birkhäuser Boston Inc., Boston, MA, 1994. Fundamental algorithms.
- [3] Alice Guionnet. Large deviations and stochastic calculus for large random matrices. *Probab. Surv.*, 1 :72–172 (electronic), 2004.
- [4] A. Khorunzhy and W. Kirsch. On asymptotic expansions and scales of spectral universality in band random matrix ensembles. *Comm. Math. Phys.*, 231(2) :223–255, 2002.
- [5] S. A. Molchanov, L. A. Pastur, and A. M. Khorunzhiĭ. Distribution of the eigenvalues of random band matrices in the limit of their infinite order. *Teoret. Mat. Fiz.*, 90(2) :163–178, 1992.